

Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems

Year: 2020 | Citations: 314 | Authors: Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, Elena L. Glassman

Abstract

Explainable artificially intelligent (XAI) systems form part of sociotechnical systems, e.g., human+AI teams tasked with making decisions. Yet, current XAI systems are rarely evaluated by measuring the performance of human+AI teams on actual decision-making tasks. We conducted two online experiments and one in-person think-aloud study to evaluate two currently common techniques for evaluating XAI systems: (1) using proxy, artificial tasks such as how well humans predict the AI's decision from the given explanations, and (2) using subjective measures of trust and preference as predictors of actual performance. The results of our experiments demonstrate that evaluations with proxy tasks did not predict the results of the evaluations with the actual decision-making tasks. Further, the subjective measures on evaluations with actual decision-making tasks did not predict the objective performance on those same tasks. Our results suggest that by employing misleading evaluation methods, our field may be inadvertently slowing its progress toward developing human+AI teams that can reliably perform better than humans or AIs alone.